# SCOPE OF HYBRID LEARNING IN INSURANCE FRAUD DEDECTION

**Dr. Ananthi Sheshasaayee[1] Surya Susan Thomas[2]**

[1]Research Supervisor, Department of Computer Science, Quaid E Millath Govt College for Women, Chennai-600 002, India.

[2]Research Scholar, Department of Computer Science, Quaid E Millath Govt College for Women, Chennai-600 002, India.

[1]ananthi.research@gmail.com, [2]susann.research@gmail.com

**Abstract-**Insurance fraud is a deliberate attempt to obtain a fraudulent upshot from an insurance claim process. This occurs when a claimant purposely attempts to obtain some benefits which may not be entitled to him/her or when an insurer knowingly denies some benefit which is due. As the awareness of getting insured has substantially increased over the years, insurance claims also increased which in turn boosted the fraudulent claims. This circumstance has led to the necessity to detect fraudulent claims at the earliest so as to reduce financial burdens. Data mining methods and techniques enhance the fraud detection in this area to a great extent. This paper discusses the role of hybrid learning data mining methods to mitigate frauds in insurance industry especially in the health insurance sectors.

**Keywords:** Insurance, Fraud detection, Data mining, KDD.

## 1. INTRODUCTION:

Insurance is a procedure by which insurance company or state assures a compensation for specified loss, damage, illness or death in return for payment of a specified premium. In other words, Insurance is a plan providing protection against a possible eventuality.

Insurance in India is broadly classified into life ,motor ,health, travel ,home ,rural, commercial and business insurance.

"The Insurance Regulatory and Development Authority", an agency of the Government of India, is the regulatory committee for the insurance sector's supervision and development in India. This organization was started in 1999 by the Government of India, for two significant reasons-to protect the interest of the policy holders and for the up gradation of the entire insurance sector right from the issue of a claim till the settlement of a claim. [1]

False insurance claims are claims filed with an intentto falsify the insurance provider. Fraudulent claims costs cores of rupees annually. These claims occur in all areas of insurance. Insurance crimes range in severity, from the less impact ones to the ones that cause cores of financial loss to the provider. [2] Fraudulent activities affect the lives of claim holders both directly and indirectly as these crimes cause insurance premiums to be higher. Insurance fraud is a significant problem, and has to be dealt with stringent action against the fraudsters.

As automation takes place in all sectors of business, fraud detection also achieved a new twist with the introduction of data mining concept. Data mining helps in detecting frauds efficiently and with more ease.

### 1.1 Data Mining:

Data Mining [3] is the process of selecting, exploring and analysing large amounts of data to mine or dig up previously unknown truth, which might prove to be a solution to the existing problem. In Insurance industry, data mining is an added advantage for the firms to gain business advantage.

Data Mining is the brain of the integrated information technology software that is usually referred as 'Business Intelligence'. These Information technology systems commences with the data warehousing followed by Online analytical processing(OLAP) and concludes with the data mining methods.[4] This paper confines to the study and analysis of the various hybrid data mining methods employed to detect health insurance frauds. It is not possible to fully eradicate frauds in this sector but can be reduced with the help of data mining methods and other businesssolutions.Figure:1 shows the data mining process for a business solution.[5]

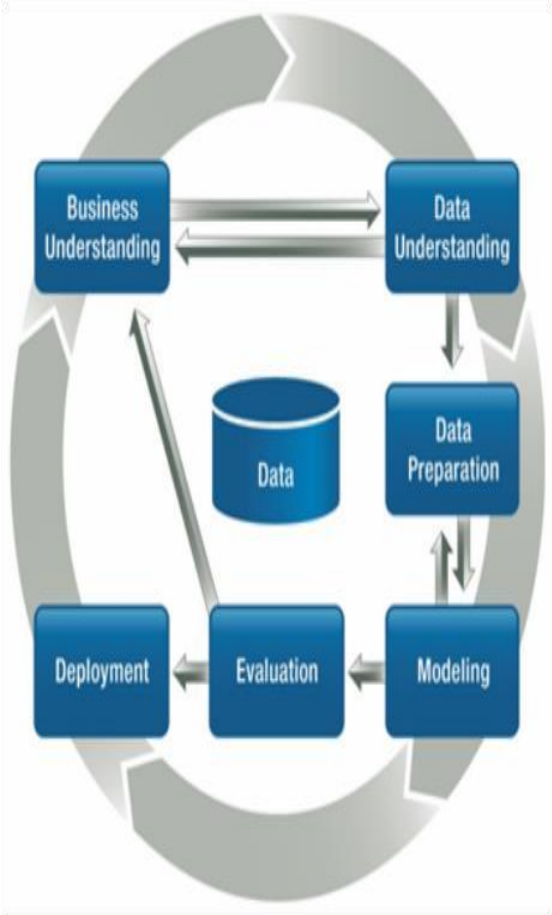


**Figure 1**

### 1.2. Detecting Frauds:

The traditional way of detecting fraud is the procedure which mainly comprises of the auditing of hundreds of claims manually. [6] But this proves to be time consuming and inefficient. This leads to the lag in the clearance of the claims which affects both the insurer and the policy holder. Automated business process and growth of computerized systems has led to sophisticated methods to detect fraud and abuse. Knowledge Discovery from Databases (KDD) emerged by combining information technology software systems and statically knowledge. Data mining is the spine of KDD.Data mining helps these insurance companies to extract useful information or knowledge from thousands of claims to form a smaller class of the doubtful claims for further scrutiny for fraud and abuse.[7]

Frauds are of different types in healthcare, mainly comprises of

- Billing services not rendered
- Up coding
- Duplicate claims
- Unbundling
- Kickbacks

### 2. SUPERVISED LEARNING METHODS:

Supervised learning methodologies works with the help of training data which includes both the input and the desired results. In some cases, correct results are identified and are given as input to the model during the learning process. Building up of a proper training validation and test set are very important.

Supervised learning is usually fast and precise.[8]. Decision tree, classification, regression are the major examples of supervised learning in insurance fraud detection domain.

### 3. UNSUPERVISED LEARNING METHODS:

Unsupervised learning is a methodology where the training set is not available and finding patterns or structure in the data has to be done independently. This method typically analyses a insurance claim's attributes in relation to other claims and discover how they are related to or differ from each other. Algorithms are ported to the devices to explore and present the interesting pattern in the data. Some known examples of this learning are k-means for clustering techniques, outlier detection and Apriori algorithm for association rule learning problems.[8]

### 4. HYBRID LEARNING METHODS:

This methodology uses mainly unlabelled and a small amount of labelled input data. The usage of some labelled data can greatly enhance the efficiency of unsupervised learning tasks.[9] The model must learn the structure to organise the data as well as make predictions. The actual problem resides in between supervised and unsupervised learning. A study followed a three-step procedure for insurance fraud detection. It applied unsupervised clustering methods on insurance claims and modelled a variety of labelled clusters. Then they used an algorithm based on a supervised classification tree and discovered rules for the allocation of each record to clusters. They generated the most effective rules for future identification of fraud behaviours. [10]

**4.1Importance of Hybrid learning:**

Labelled data are difficult to obtain while unlabelled data are plenty, therefore hybrid learning is a good solution to minimize human effort and improve accuracy. [11]

The aim of hybrid learning is to explore the combination of labelled and unlabelled data which might change the learning environment and modulate algorithms that take advantage of such a combination. This learning is interesting in machine learning and data mining concepts because it can make use of the unlabelled data to improvise supervised learning which are limited and expensive. Semi-supervised learning models such as self-training mixture models-training and Multiview learning, graph-based methods and semi-supervised support vector machines are the existing ones.

**4.2. Methods in Hybrid learning:**

Because of the scarcity of the supervised or labelled data, hybrid learning models make strong model predictions. Soone should carefully choose a model which best fits the problem structure.

**4.2.1 Generative Models**

It is one of the oldest hybrid learning methods. This model is used for randomly generating observable data values, typically data with hidden parameters.

**4.2.2 Graph based models**

Graph-based hybrid methods are a representation where the nodes are trained and untrained data in the dataset, and the edges reflect the similarity of the data. This method assumes label smoothness in the graph. Graph methods are trans dative in nature[12]

**4.2.3 Self-Training**

In self-training, a classifier is first trained with labelled data and then classified with the unlabelled data again. The most confident unlabelled points, are then predicted with the help of labelled points. The classifier is retrained and the procedure is repeated.

**5. RELATED WORKS:**

A number of researchers have developed hybrid methods by combining supervised and unsupervised methods. Unsupervised method proceeded by supervised method usually give way to discovery of knowledge in a hierarchical manner.

Williams and Huang combined clustering algorithms and decision trees to detect insurance subscribers' fraud. A three-step "divide and conquer" procedure was used to find the solution. [13]

Major et al used hybrid learning in EFD (Electronic Fraud Detection). They utilized knowledge discovery techniques on two levels. First, integrating knowledge with statistical information assessment and second machine learning was used to develop new rules and improve the identification process. [14]

Abraham et al introduced two hybrid approaches for modelling "Intrusion Detection System(IDS)".Decision trees and SVMs are combined as a top-to-bottom hybrid model and a novel approach joining the base classifiers. The hybrid abuse detection model combines the classifiers and other hybrid machine learning paradigms to maximize detection accuracy and curtail complexity in computation. [15].

Zhang et al overviewed the advanced supervised machine learning and natural language processing techniques to the problem of detecting anomalies in financial reporting documents.[16].

Peng et al constructed a model to detect suspicious health care frauds from large databases using clustering techniques. They applied two clustering methods namely SAS EM and CLUTO to health insurance dataset and compares their capability.CLUTO was faster but SAS EM gave more useful clusters. [17].

Anuradha et al integrates SVM(Support Vector Machine) and ECM(Evolving Clustering Method) in health insurance field for fraud detection. They used SVM algorithm was used for classification and EVM algorithm was used for clustering. Applying these two methods, the system was trained to draw a boundary between legitimate and fraudulent claims with more accuracy.[18].

Rashidian et al applied supervised and unsupervised data mining techniques to detect healthcare frauds. SVM, neural networks, genetic algorithms, decision trees were used from supervised learning and clustering, outlier detection and association rules were used from the unsupervised learning methodology. They were able to retrieve better results by the hybrid learning method than by using any one methodology. [19].

Kose et al developed a novel framework to detect fraudulent cases independent of the players and items

involved in the abuse. Interactive machine learning that allows incorporating expert knowledge in an unsupervised setting was employed. They used pairwise comparison method of Analytical Hierarchical Processing (AHP) for weighing the actors and properties and a maximization method for clustering similar players, two stage data warehousing for proactive risk calculations, visualization tools for effective analysing and Z-score and standardization to calculate the risks. The framework which was named eFAD suite effectively handles the fragmented nature of the abusive behaviours. [20]

There are three steps formed by hot spot methodology [21], they are; i) k-means clustering algorithm for cluster detection is used because only this algorithm is only very cheap and which is used for very large data set others are very expensive, ii) The result of C4.5 algorithm decision tree could be convert into a set of rules and ordered and iii) building the statistical summaries of the entities are associated with each rule by visualization tools for rule evaluation.

The model for credit fraud [22] suggests a classification technique with fraud/valid attribute, and a cluster followed by a classification technique without fraud/legal attributes. Kohonen's Self-Organizing Feature Map [23] is used for categorize automobile injuries are claimed depending upon a type of fraud.

Classification technique has proved that it will be a very effective in fraud detection [24], and therefore it can be applied to categorize the crime data. The data mining models of distributed ways are used as a realistic cost model to estimate C4.5, CART and Naïve Bayes classification model. This model is useful for credit card transaction. The neural data mining method is used as rule based association rule to mine representative data. This approach tells us how importance to use non-numeric data in fraud detection.

SAS Enterprise miner Software [25] is depending on associated rules, cluster detection and classification technique to detect fraud claims. The Bayesian Belief Network (BBN) and Artificial Neural Network (ANN) studies are used for the STAE algorithm for BBN in fraud detection and back propagation for ANN[26]. The results shows that BBNs are faster to train, but it will be very slow when it is applied to new instances. The ASPECT group [27]focused on neural networks to train the current users profile history. The current profile of a caller's and a profile history are to be compared to find probable fraud. Building an adaptive fraud detection frame works are related by an event-driven approach which is assigning for fraud scores to detect the fraud. The frame work detects a fraud using rules, BBN is called as a Mass Detection tool to detect fake claims which is used as a rule generator called Suspicious Building Tool.

Internal fraud detection has a determining false financial reporting by management [28] and irregular retail transaction used by employees.

There are four types of insurance fraud detection method, they are;

- Home insurance
- Crop insurance
- Automobile insurance fraud detection and
- Health insurance

Single Meta classifiers are used for selecting the best base classifier and then it is combined with these base classifier predictions to improve the rate savers. Fraud detection of Credit card is referred as screening credit application and /or logged credit card transaction. Online seller and buyers [29, 30] are been monitored by automated systems. Here in fraud detection government organizations such as tax and customs [31, 32] are also been reported.

Most of the fraud detection systems is used by either supervised or unsupervised learning algorithm. The implemented Supervised learning algorithms are Support Vector Machine, Neural Networks Logistic regression. There is no single algorithm between them which will give a satisfactory result. Supervised Learning algorithms results as fail to detect unexpected conditions which could be handle if unsupervised learning algorithm is conducted.

There are three parties which will be involved in the commission of health care fraud, they are;

- Service providers, including doctors, hospitals, ambulance companies and laboratories
- Insurance subscribers, it also includes patients and patients 'employers and
- Insurance carriers

Who will receive a regular payment from their subscribers and pay health care cost on behalf of their subscribers it it also included for government and private health department and insurance companies.

We can classify the fraud behaviours according which party is committing the fraud work.

1. Service Providers Fraud:
   - Billing Service which will not be performed actual.
   - Unbundling: i.e., billing has each stage of procedure which has a separate treatment.
   - Up coding: i.e., billing in an expensive way.
   - Performing unnecessary medical treatment for claiming the health insurance and
   - Misrepresenting non-covered treatment as medical treatment for claiming as a medical insurance
2. Insurance subscribers' fraud:
   - The false record of an employment/ eligible for getting a lower premium
   - Falsely claims as a medical service which is not received and
   - Using others record to claim the insurance money
3. Insurance carriers' fraud:
   - False Compensations
   - False service/ benefit statement.

Among these three types of fraud, only one is committed by service provider's account for the great proportion for the total health care fraud and abuse. While the huge majority of service providers are honest and moral, but very few are dishonest one might have various possible way to commit fraud on a broad scale which damage to the health care system [32]. Some service providers are fraud, where they are involving medical transportation. Therefore service providers fraud is more urgent problem to secure more of quality and safety of a health care system.

In Health Insurance Company fraud is speeded widely. Usually, health insurance companies will be using experiential rules to detect frauds. These rules are summarized from earlier fraud case which is used for detecting the fraud by both ways, one is through human inspection and another one is interaction with external person. The increasing of database the usual way of fraud detection becomes failure. The shocking news is that fraud is keep increasing every year in health insurance. So it is very needful to find a new advance approach to detect suspicious health care frauds from large database.

The raw data for health care fraud detection arise mostly from insurance carriers; it also includes government and private health department and insurance companies. Major government health departments are being reported in the literature including US Health Care Financing Administration (HCFA)[33,34], the Bureau of National Health Insurance (NHI) in Taiwan [35,36,37,38,39,] and the Health Insurance Commission (HIC) in Australia [40,41,42,43,44]

## 6. CONCLUSION:

Supervised learning methods applied to detect frauds gave accurate results but to access labelled data is expensive and scarce. In order to overcome this, unsupervised learning methods were introduced and this too had limitations of training the data which was time consuming and results might not be accurate as the supervised learning. With the induction of the hybrid learning method, fraud detection in insurance industry has become easier and cost-effective. It gives an insight to improvise the existing fraud detection methodologies not only in insurance sector but also in other business sectors.

## REFERENCE:

1. IRDAI;" History of Insurance in India", Ref: IRDA/GEN/06/2007, July 2007.
2. Ngufor, Che, and A. Wojtusiak. "Unsupervised labeling of data for supervised learning and its application to medical claims prediction." Computer Science14 (2013).
3. R. Karpagam, Dr. S. Suganya, "Applications of data mining and algorithms in education – a survey", International Journal of Innovations in Scientific and Engineering Research (IJISER), Vol 3 Issue 4, APR 2016.
4. Joudaki, Hossein, et al. "Using data mining to detect health care fraud and abuse: a review of literature." Global journal of health science 7.1 (2014): 194-202
5. http://www.rosebt.com/blog/category/enterprise%20data%20warehousing%20platforms/2
6. Ortega, Pedro A., Cristián J. Figueroa, and Gonzalo A. Ruz. "A Medical Claim Fraud/Abuse Detection System based on Data Mining: A Case Study in Chile." DMIN 6 (2006): 26-29.
7. Faseela, V. S., and P. Thangam. "A Review on Health Insurance Claim Fraud Detection." International Journal of Engineering Research Science (IJOER) 1 (2015).
8. http://www.astro.caltech.edu/~george/aybi199/Donalek_Classif.pdf
9. https://www.nesta.org.uk/sites/default/files/machines_that_learn_in_the_wild.pdf
10. Major, John A., and Dan R. Riedinger. "EFD: A Hybrid Knowledge/Statistical-Based System for the Detection of Fraud." Journal of Risk and Insurance69.3 (2002): 309-324.

11. PunamDevidasBagul, Sachin Bojewar, Ankit Sanghavi." Survey on Hybrid Approach for Fraud Detection in Health Insurance". International Journal of Innovative Research in Computer and Communication EngineeringVol4, Issue 4, April (2016).
12. Zhu, Xiaojin. "Semi-supervised learning literature survey." (2005).
13. Williams, Graham J., and Zhexue Huang."Mining the knowledge mine." Australian Joint Conference on Artificial Intelligence. Springer Berlin Heidelberg, 1997.
14. Major, John A., and Dan R. Riedinger. "EFD: A Hybrid Knowledge/Statistical-Based System for the Detection of Fraud." Journal of Risk and Insurance69.3 (2002): 309-324
15. Peddabachigari, Sandhya, et al. "Modeling intrusion detection system using hybrid intelligent systems." Journal of network and computer applications30.1 (2007): 114-132.
16. Seemakurthi, Prasad, Shuhao Zhang, and Yibing Qi. "Detection of fraudulent financial reports with machine learning techniques." Systems and Information Engineering Design Symposium (SIEDS), 2015. IEEE, 2015.
17. Peng, Yi, et al. "Application of clustering methods to health insurance fraud detection." 2006 International Conference on Service Systems and Service Management. Vol. 1. IEEE, 2006.
18. Rawte, Vipula, and G. Anuradha. "Fraud detection in health insurance using data mining techniques." Communication, Information & Computing Technology (ICCICT), 2015 International Conference on. IEEE, 2015.
19. Joudaki, Hossein, et al. "Using data mining to detect health care fraud and abuse: a review of literature." Global journal of health science 7.1 (2014): 194-202.
20. Kose, Ilker, Mehmet Gokturk, and Kemal Kilic. "An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance." Applied Soft Computing 36 (2015): 28329.
21. Williams, G.: Evolutionary Hot Spots Data Mining: An Architecture for Exploring for Interesting Discoveries. In: 3rd Pacific-Asia Conference in Knowledge Discovery and Data Mining, Beijing, China(1999).
22. Groth, R.: Data Mining: A Hands-on Approach for Business Professionals, Prentice Hall, pp. 209-212(1998).
23. Brockett, P., Derrig, R., Golden, L., Levine, A. & Alpert, M.: Fraud Classification using Principal Component Analysis of RIDITs. Journal of Risk and Insurance 69(3): 341-371(2002).
24. Chen, R., Chiu, M., Huang, Y., Chen, L.: Detecting Credit Card Fraud by Using Questionnaire-Responded Transaction Model Based on Support Vector Machines. In: IDEAL2004, 800--806(2004).
25. SAS, e-Intelligence Data Mining in the Insurance industry: Solving Business problems using SAS Enterprise Miner Software. White Paper(2000).
26. Maes, S., Tuyls, K., Vanschoenwinkel, B. & Manderick, B.: Credit Card Fraud Detection using Bayesian and Neural Networks. Proc. of the 1st International NAISO Congress on Neuro Fuzzy Technologies (2002).
27. Weatherford, M.: Mining for Fraud. In: IEEE Intelligent Systems (2002).
28. Lin, J., Hwang, M., Becker, J.: A Fuzzy Neural Network for Assessing the Risk of Fraudulent Financial Reporting. J. of Managerial Auditing, 18(8), 657--665(2003).
29. Bhargava, B., Zhong, Y., Lu, Y.: Fraud Formalization and Detection. In: DaWaK2003, 330--339(2003).
30. Sherman, E.: Fighting Web Fraud. Newsweek, June 10(2002).
31. Bonchi, F., Giannotti, F., Mainetto, G., Pedreschi, D.: A Classification-based Methodology for Planning Auditing Strategies in Fraud Detection. In: SIGKDD99, 175--184(1999).
32. Shao, H., Zhao, H., Chang, G.: Applying Data Mining to Detect Fraud Behavior in Customs Declaration. In: 1st International Conference on Machine Learning and Cybernetics, 1241--1244(2002).
33. GAO (1996) Health Care Fraud: Information-Sharing Proposals to Improve Enforcement Effects. Report of United States General Accounting Office
34. Shapiro AF (2002) The merging of neural networks, fuzzy logic, and genetic algorithms. Insurance: Mathematics and Economics 31:115–131
35. Chan CL, Lan CH (2001) A data mining technique combining fuzzy sets theory and Bayesian classifier—an application of auditing the health insurance fee. In Proceedings of the International Conference on Artificial Intelligence, 402–408
36. Hwang SY, Wei CP, Yang WS (2003) Discovery of temporal patterns from process instances. Comp Ind 53:345–364.
37. Wei CP, Hwang SY, Yang WS (2000) Mining frequent temporal patterns in process databases. Proceedings of international workshop on information technologies and systems, Australia, 175–180.
38. Yang WS, Hwang SY (2006) A process-mining framework for the detection of healthcare fraud and abuse. Expert Syst Appl 31:56–68.
39. Yang WS (2003) A Process Pattern Mining Framework for the Detection of Health Care Fraud and Abuse, Ph.D. thesis, National Sun Yat-Sen University, Taiwan
40. He H, Hawkins S, Graco W, Yao X (2000) Application of Genetic Algorithms and k-Nearest Neighbour method in real world medical fraud detection problem. Journal of Advanced Computational Intelligence and Intelligent Informatics 4(2):130–137
41. He H, Wang J, Graco W, Hawkins S (1997) Application of neural networks to detection of medical fraud. Expert Syst Appl 13:329–336
42. Hubick KT (1992) artificial neural networks in Australia. Department of Industry, Technology and Commerce, CPN Publications, Canberra.

43. Williams G, Huang Z (1997) Mining the knowledge mine: The Hot Spots methodology for mining large real world databases. Lect Notes Comput Sci 1342:340–348
44. Yamanishi K, Takeuchi J, Williams G, Milne P (2004) On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. Data Mining and Knowledge Discovery 8:275–300.